A review of post-study and post-task subjective questionnaires to guide assessment of system usability

Hodrien, A and Fernando, TP

Journal of Usability Studies
Vol. 16, Issue 3, May 2021, pp. 203–232

A Review of Post-Study and Post-Task Subjective Questionnaires to Guide Assessment of System Usability

Andrew Hodrien
University of Salford
Maxwell Building
43 Crescent
Salford, Greater
Manchester
UK
a.hodrien1@salford.ac.uk

Terrence Fernando
University of Salford
Maxwell Building
43 Crescent
Salford, Greater
Manchester
UK
t.fernando@salford.ac.uk

Abstract

Usability is a key consideration when developing an interactive software application because of the various outcomes it can produce. Accordingly, numerous evaluation methods have been proposed, however, a recent review of usability methods concluded there is no current consensus on models applied to usability. While questionnaires are a commonly used measure, it is unclear which questionnaire(s) are most appropriate for a given context, so new usability researchers face confusion over which to select. To aid questionnaire selection, the appropriate type (post-study or post-task), general structure and content, specific advantages and disadvantages, aspects of usability covered, and psychometric quality should be considered. This paper presents a literature review and analysis of general post-study and post-task usability measures. Questionnaires are weighed up and discussed on each aspect, so practitioners can gain a holistic overview and identify relative strengths of individual questionnaires within their questionnaire type. Overall recommendations and specific questionnaire suggestions are provided for guidance, along with how future research can expand the review.

Keywords

questionnaire, user experience, post-study, post-task, aspects of usability, psychometrics

Introduction

Usability applied to a system can broadly be defined as how easy it is to use the system (Sagar & Saha, 2017). More specifically, standard guidelines defining usability (ISO, 1998) highlight that usability cannot be applied to a specific aspect of a system; instead, a usable system is one within an appropriate context, including the task, the user's experience, and environment (Brooke, 2013). Usability is of key importance as, when it is limited, it can result in system failure and dissatisfied, unproductive users (Madan & Dubey, 2012). In some situations, such a lack of usability could have even more serious implications.

Currently, numerous usability evaluation methods have been proposed (Madan & Dubey, 2012), but which of these is most appropriate is unclear (Lewis, 1995). Sagar and Saha's (2017) recent review concluded that developers generally do not have enough knowledge when deciding the most appropriate method for the context. Despite this, the effort and expenditure used for data collection and analysis often encourage the use of subjective measures (Longo, 2018). Accordingly, Sagar and Saha identified that questionnaires are frequently used, while arguing a need for a single measure covering all aspects of usability. Until this is constructed, new usability researchers potentially face confusion over which questionnaires to select, and multiple measures may be needed to cover all aspects. Sauro and Lewis (2012) previously detailed numerous questionnaires, but highlighted that there has been limited research on directly comparing the different questionnaires. Therefore, there is a need for a holistic overview in weighing up questionnaires. However, in order to decide the most appropriate measures, various factors should be considered as outlined in following sections.

Type of Questionnaire, General Structure/Content, and Advantages and Disadvantages

Firstly, appropriateness of the questionnaire to the usability study should be considered. Sauro and Lewis (2012) categorized questionnaires broadly as "post-study," "post-task," "website," and "other," thus if a general usability measure is required for hardware or software applications, the first two categories are most suitable. Post-study and post-task questionnaires differentiate the focus points in a usability study: presented once at the end of a study or presented straight after a task/scenario, respectively. Next, questionnaire structure and content should be considered based on the needs of the study. Lastly, questionnaires' advantages and disadvantages may have potential relevance for a specific usability study, thus should be considered when choosing an individual questionnaire and/or when comparing between specific questionnaires.

Aspects of Usability

With Sagar and Saha (2017) noting that a single measure does not cover all aspects of usability, it is potentially unclear what each measure does cover. In addition, they found no current consensus on the models applied to usability, possibly due to its ambiguous nature (Madan & Dubey, 2012). Frøkjær et al. (2000) highlighted that there is considerable confusion regarding usability, with practitioners adopting either a narrow or broad definition. The former is risky, as Brooke (2013) elaborated: A system allowing task completion (effectiveness) at the expense of time and effort (efficiency) and felt to be unsatisfactory (satisfaction), could not be described as usable.

While effectiveness and efficiency are commonly considered by usability researchers as objective usability components (Lewis et al., 2015), users may also consider how effective or efficient the system is, thus also bringing perceptual aspects. Furthermore, the relative importance of usability aspects may depend on the task complexity. Frøkjær et al. (2000) noted that routine tasks require an efficient execution of actions in a sequence, thus overall usability may be determined from subjective efficiency/task completion time; whereas, more complex tasks have no predetermined route for task completion. This means performance is largely based on the user identifying a solution for the task, with the efficient execution of actions being of less importance. Such differences might determine whether efficiency or effectiveness takes dominance in assessing usability.

Frøkjær et al. (2000) noted that while some researchers assume relationships between aspects of usability, such relationships depend on numerous factors (e.g., task complexity dictating the relative importance of efficiency or effectiveness). Frøkjær et al. reviewed several studies,

finding that only 8 out of the 19 covered three main aspects of usability, with 11 assuming relationships between the aspects. In addition, from their own experimental evidence, these authors found only a negligible correlation between effectiveness and efficiency. Relationship assumptions are potentially risky, as this can result in selecting a narrow range of usability measures, with one or more aspect(s) of usability being ignored (Frøkjær et al., 2000). They concluded that, due to limited understanding of relationships, aspects should be considered as independent and all should be included, along with considering the context in order to identify the critical measures.

Psychometric Quality

The psychometric properties of a questionnaire should be considered. As with questionnaire development, it is standard to report on a questionnaire's psychometric qualification (i.e., its reliability, validity, and sensitivity) to support its objectivity to verify measurements, replicate studies, and provide scientific generalization (Sauro & Lewis, 2012). There are a limited number of studies that have directly compared questionnaires on their psychometric performance (e.g., see Lewis, 1995; Sauro & Dumas, 2009; Tedesco & Tullis, 2006; Tullis & Stetson, 2004), thus comparisons across studies may be necessary.

Summary

In deciding the most suitable questionnaire(s) for a usability study, various factors that have been mentioned above should be considered: the study aims, appropriateness for the system, type of questionnaire, structure and content, advantages and disadvantages, aspects of usability covered, and psychometric quality. The purpose of this review paper is to provide a roadmap for decisions on questionnaire selection for usability studies. The review aims to weigh up questionnaires on specific areas to ultimately gain a holistic overview and to highlight various factors to consider in the decision-making process.

Questionnaire Literature Review Method

To identify relevant questionnaires, the details for each, and understand the overall field of usability, a literature review was conducted (via Google Scholar) using a "snowball sampling" type approach and paying attention to specific articles. Snowballing involves exploring the references' lists and paper citations, referred to as backward and forward snowballing, respectively (Wohlin, 2014). Following advice from Webster and Watson (2002), a structured approach identifying key papers and citations ensures a relatively complete number of appropriate sources. As the aims of this study were to identify general usability questionnaires and the main usability concepts, this was considered suitable. Evidence of the review nearing completion is when new concepts (i.e., questionnaires) cease to emerge (Webster & Watson, 2002). General articles discussing usability, focusing on a specific measure or including psychometric data for a specific questionnaire, were explored, rather than articles focusing on the usability of a specific system/technology. Potential questionnaires were reviewed based on the following inclusion/exclusion criteria.

The questionnaires included in this research were the ones that assess some form of system or technology (hardware or software), were general in nature (e.g., assessing computer systems), have at least some psychometric data available, are the latest versions unless multiple versions exist for different scenarios (e.g., one designed for usability experiments and another for field testing), and focus on assessing usability (where unclear, the prevalence of subscales or items focuses on usability, e.g., at least over 50% of the questionnaire content). The questionnaires excluded from this research were the ones that assess usability of something other than a system or piece of technology, were for a specific system/technology or specifically designed for assessing websites, have no available psychometric data, have been improved upon and replaced, or do not focus on usability.

Structure of Paper

Questionnaires were categorized as either post-study questionnaire, presented once at the end of a usability study, or post-task, presented straight after a task or scenario. This paper is split between these broad questionnaire types. Each section details analyses in tables on structure and content of questionnaires, advantages and disadvantages, aspects of usability covered by questionnaires, and psychometric quality of questionnaires, including how the table is organized, information (where relevant) on specific questionnaires preceding the table, and a discussion/summary following the table. Details of the sub-sections are discussed in the following sections.

Structure and Content of Questionnaires

The purpose of this section is to review the general details of questionnaires for familiarization, including identifying structural and content aspects that may support deciding the most appropriate measure(s) to use. For each, its structure, type, number of points, and number of items were identified, along with any specific background information. Questionnaire structure involves whether specific items are calculated together to measure a component of usability (i.e., a subscale) or whether all items produce an overall score. Questionnaires measure usability utilizing either one or both ways, which may be of use for questionnaire choice.

Advantages and Disadvantages of Questionnaires

The purpose of this section is to review the specific advantages or disadvantages of each questionnaire, which may or may not have relevance for a usability study, to help inform decisions on questionnaire(s) selection.

Aspects of Usability Covered by Questionnaires

The purpose of this section is to review which usability aspects are covered by each questionnaire. This would be beneficial for usability researchers to help decide the most appropriate measure(s) to include in their usability assessments. Sagar and Saha's (2017) review highlighted effectiveness, efficiency, satisfaction, and learnability as common assessments of usability; therefore, these aspects are focused on in the review below. In order to avoid ambiguity, the following definitions are used for the usability elements:

- Effectiveness: This includes the ability of the user to complete the relevant task or whether they achieved the correct outcome or conclusion.
- Efficiency: This includes the amount of resource expenditure involved in completing a task, covered by how quickly a task can be completed, how much effort is required, or how many mistakes are made.
- Satisfaction: This includes how much comfort is experienced in using a system, how positive the user feels toward the system, or how satisfied the user is with the general usability of the system.
- Learnability: This includes how easily or quickly the system can be learned.

A separate, but potentially related, construct to usability is "workload" (Longo, 2018), which is the user effort in achieving their task. Hart (2006) noted the relevance of workload to system performance, and thus usability. However, as Longo highlighted, the relationship between these is underexplored and that the constructs are not correlated overall. This suggests workload might not be considered an aspect of usability. Additionally, there is a risk of confusion with overlap between efficiency and workload, as errors made could be due to both limited system efficiency and greater workload. Also, some workload aspects—effort required, frustration experienced, and performance—overlap with efficiency, satisfaction, and effectiveness, respectively. For these reasons, workload is not included here.

Psychometric Quality of Questionnaires

As mentioned in Sauro and Lewis (2012), the psychometric properties of a questionnaire (reliability, validity, and sensitivity) support its objectivity to verify measurements, replicate studies, and provide scientific generalization. The purpose of this section is to review the psychometric support for each questionnaire to help inform decisions on questionnaire(s) selection. Firstly, a brief introduction to the psychometric properties is presented in the following sections for completeness.

Reliability

Reliability is how consistent a questionnaire is, often measured using Cronbach's alpha/coefficient alpha (Lewis, 1995); the average of all the coefficients from every possible combination of items split into two (Gliem & Gliem, 2003). Other ways include split-half reliability and test-retest reliability (Sauro & Lewis, 2012), where half the items correlated with the other half and scores correlated with those at a later time, respectively.

Validity

Validity is how much a questionnaire actually measures what it intends to (Lewis, 1995). Content validity is where items are representative of the construct, and factorial validity is where factor analysis has identified the structure of the scale. Criterion validity, the relationship between the scale and another expected to correlate with it, is often measured with the Pearson correlation coefficient (Lewis, 1995).

Sensitivity

Sensitivity is the scale's ability to detect relevant differences, as a reliable and valid scale should be sensitive to such aspects (Lewis, 2002). It is normally calculated by looking for statistical differences between things that are expected to result in different usability scores. For example, the measure should detect differences between systems, tasks of variable difficulty, and the expertise of the users.

Psychometric Data Collection Method

Any relevant psychometric data identified for each questionnaire in the initial literature review was included in the following analysis sections. If no information was identified for a specific type of psychometric data (e.g., reliability) then this was sought via a keyword search in Google Scholar using the questionnaire name and the type of psychometric data missing (i.e., reliability, validity, or sensitivity).

Post-Study Questionnaires Analysis

This section includes those questionnaires that were identified as being a post-study questionnaire type.

Structure and Content of Post-Study Questionnaires

Table 1 summarizes the key post-study questionnaires used for assessing usability, ordered alphabetically by questionnaire name, as at this stage no questionnaire takes precedence over another.

Table 1. Summary of Post-Study Questionnaires with General Details

Questionnaire (acronym)	Questionnaire general details			
	Subscales	**Type of scale**	**Number of points**	**Number of items**
AttrakDiff 2	Hedonic quality-identification (HQI), Hedonic quality-stimulation (HQS), Pragmatic quality (PQ), Attraction (ATT)[ab]	SDS	7	28
Computer System Usability Questionnaire (CSUQ)	System Usefulness, Information Quality, and Interface Quality[a]	LS	7	19
End User Computer Satisfaction (EUCS)	Content, Accuracy, Format, Ease of use, and Timeliness[a]	LS	5	12
Modular Evaluation of Components of User Experience (meCUE)	Usefulness, Usability, Visual aesthetics, Social identity: Status, Social identity: Commitment, Positive emotions, Negative emotions[a]	LS (subscales), SDS (overall evaluation)	7 (subscales), 21 (overall evaluation)	34
Post-Study System Usability Questionnaire (PSSUQ)	System Usefulness, Information Quality, and Interface Quality[a]	LS	7	19
Purdue Usability Questionnaire (PUTQ)	N/A	SDS	7	100
Questionnaire for User Interface Satisfaction (QUIS)	Screen, Terminology and System information, Learning, and System capabilities[a]	SDS	10	27
Software Usability Measurement Inventory (SUMI)	Efficiency, Affect, Helpfulness, Control, and Learnability[a]	LS	3	50 (10 per subscale, 25 for global scale)
System Usability Scale (SUS)	Usability and Learnability[a]	LS	5	10
Technology Acceptance Model (TAM)	Ease of use and Usefulness	LS	7	12
User Experience Questionnaire (UEQ)	Perspicuity, Efficiency, Dependability, Novelty, and Stimulation[a]	SDS	7	26
Usability Metric for User Experience (UMUX)	N/A	LS	7	4
Usability Metric for User Experience-LITE (UMUX-LITE)	N/A	LS	7	2
Usefulness, Satisfaction, and Ease of Use (USE)	Usefulness, Ease of use, Ease of learning, and Satisfaction	LS	7	30

Note. Unless mentioned, items refer to the total number of items, not number per subscale. N/A = Not applicable (The questionnaire has no subscales; it just includes an overall score). SDS = Semantic differential scale. LS = Likert scale.

[a] Along with subscales, the questionnaire includes an overall score. [b] While the AttrakDiff 2 does not have a traditional overall score (e.g., averaged from subscales), the ATT subscale measures the overall attraction of a system, so it is considered an overall score here.

The SUS (available in Sauro & Lewis, 2012) is a widely used measure of general usability (Brooke, 2013); however, some doubt has been cast on its structure (Lewis & Sauro, 2017; Lewis et al., 2015), suggesting that the scale should be unidimensional. Lewis et al. (2015, see reference for questionnaire) developed an alternative version of the SUS, AltUsability, similar in style but the items cover issues involving navigation, findability, familiarity, efficiency, a feeling of control, and visual appeal. While this was not fully developed into a standardized questionnaire (thus not included here as a separate measure), it is mentioned here for completeness and for those interested in the potential use of alternative items. Similarly, the UMUX (available in Finstad, 2010) was based on the SUS but is generally used for situations where a shorter scale is required due to time restraints (Lewis et al., 2015). From the UMUX, Lewis et al. (2013; see reference for questionnaire) developed an ultrashort measure, the UMUX-LITE with items chosen based on their connection to the TAM. The TAM (available in Davis, 1989) is an influential questionnaire from the market research field; it was inspired by perceptions that being easy to use and usefulness are the main factors influencing the use of technology (Lewis et al., 2015).

The QUIS was originally published in a book by Shneiderman (1987); however, it has since been developed through numerous versions, along with short and long forms of the scale, including space to provide feedback about the system (Kirakowski, 1994). It was created due to limited questionnaires being available that exclusively focused on how an interface is evaluated (Chin, Diehl, & Norman, 1988; see reference for questionnaire). The SUMI (available in Kirakowski, n.d.) replaced the Computer User Satisfaction Inventory (CUSI) and measures perception of software quality (i.e., user satisfaction; Kirakowski, 1994). It was developed from studies exploring the SUS, QUIS, and specifically the CUSI, and the state of the art, only allowing comparisons between systems, rather than absolute benchmarks (Kirakowski, 1994).

The EUCS was developed to be a standardized measure of a user's satisfaction with a computer system, based on satisfaction leading to system use rather than the other way around (Doll & Torkzadeh, 1988; see reference for questionnaire). The PSSUQ (available in Lewis, 1992) assesses subjective satisfaction of a computer system, originally measuring performance, usability issues, and the satisfaction of the user (Lewis, 2002). The wording of the questions was altered to create a version (CSUQ, available in Lewis, 1995) suitable for field settings or surveys, instead of a usability evaluation focusing on scenarios (Lewis, 2002).

The USE questionnaire was developed to gain information applicable across domains (available in Lund, 2001). Sauro and Lewis (2012) noted that while traditional psychometric techniques were used in its development, the data on such techniques were not published. However, psychometric testing has been recently completed (Gao et al., 2018) enabling the USE to be included in this paper. The PUTQ aims to measure effectiveness, efficiency, and satisfaction of a system (available in Lin et al., 1997), with items based on eight areas of human-computer interaction identified from the theory of information processing: compatibility, consistency, flexibility, learnability, minimal action, minimal memory load, perceptual limitation, and user guidance. While being a sizable questionnaire, only relevant items are completed, with the presence or absence of these being assessed first, then rated from worst case to best case, along with weighting each item's relative importance.

Three questionnaires emphasized emotional consequences of system use and distinguished between classical/emotional usability and pragmatic/hedonic usability, respectively (Lewis & Sauro, 2020). The AttrakDiff 2 (http://attrakdiff.de/index-en.html) is the latest version of a questionnaire developed by Hassenzahl et al. (2003); it is based on a model assuming that pragmatic and hedonic product attributes emerge from combining product features with the user's expectations (Hassenzahl, 2004). Hassenzahl (2004) further distinguished between different types of hedonic attributes as stimulation and identification. Similarly, the UEQ (https://www.ueq-online.org/) differentiates between pragmatic and hedonic qualities but further distinguishes between perspicuity, efficiency, and dependability aspects of pragmatic quality, and hedonic quality was comprised of novelty and stimulation aspects, along with measuring overall attractiveness (Laugwitz et al., 2008). A short 8-item version has also been developed (see Schrepp et al., 2017). Lastly, the meCUE (http://mecue.de/english/home.html) was based on the Components model of User Experience (Thüring & Mahlke, 2007) that

distinguishes between instrumental (usefulness and usability) and non-instrumental (aesthetics, status, and commitment) qualities, emotions, and consequences (intention to use; Minge et al., 2017). As consequences of usage are not strictly usability, this subscale was not included in the tables.

Benedek and Miner (2002) highlighted that, in usability studies, it is difficult to gain input on intangible aspects, such as the "desirability" of a system, largely due to the limitation of Likert scales' content having less meaning to a participant. Additionally, there is a risk of very similar and usually positive ratings for questions. While mentioning interviews are useful, Benedek and Miner noted the time-consuming nature of these, the difficulties in eliciting honest or negative feedback, and the challenge of data analysis. For these reasons, Microsoft's Product Reaction Cards (available in Benedek & Miner, 2002) were developed, involving a card sorting exercise where participants choose freely from cards with words describing their experience and provide feedback on their choices. An alternative version, Words (available in Tullis & Stetson, 2004), takes the form of a questionnaire with individual words and check boxes to select from, along with these words being secretly classified as being either "positive" or "negative" in nature. This enables the calculation of an overall score from the percentage of the positive words out of the total number selected. Despite this, the measure is more qualitative in nature so it was not included as one of the questionnaires under review, but mentioned here for completeness.

Advantages and Disadvantages of Post-Study Questionnaires

Table 2 presents a summary of the advantages and disadvantages of each respective questionnaire. Questionnaires are ordered alphabetically as while some have more advantages and less disadvantages than others; the relevance or importance of each is subject to the individual usability study.

Table 2. Summary of Advantages and Disadvantages of Post-Study Questionnaires

Questionnaire	Advantages	Disadvantages
AttrakDiff 2	• Free to use • Covers a wide range of areas including pragmatic and hedonic aspects	• Quite time-consuming to complete
CSUQ	• Free to use • Generalizable • Somewhat flexible, with the option to add items and three specific items can be removed if desired • Normative data available	N/A
EUCS	• Free to use • Quick to complete • Generalizable	N/A
meCUE	• Free to use • Covers a wide range of areas including pragmatic and hedonic aspects	• Quite time-consuming to complete
PSSUQ	• Free to use • Generalizable • Somewhat flexible, with the option to add items and three specific items can be removed if desired • Normative data available	N/A
PUTQ	• Free to use • Covers a large range of areas	• Items focus on traditional graphical user interface software (i.e., visual display, keyboard,

Questionnaire	Advantages	Disadvantages
		mouse etc.), so not generalizable to other types of systems • Very time-consuming to complete
QUIS	N/A	• Requires a license fee • Issues have been raised for the Screen and Terminology and System Information subscales (see Kirakowski, 1994)
SUMI	• Applicable for wide range of systems and has been utilized variously by companies • Quick to complete • Can use with small samples (10-12) • Report provided with data calculations • Normative data available	• Requires a license fee • Lack of control with data being calculated externally and researchers being sent a data report • Time-consuming to complete
SUS	• Free to use • Quick to complete • Alternating or positive tone of items • Flexibility of changing the item wording • Particularly reliable with small sample sizes (Tullis & Stetson, 2004) • Generalizable across a wide range of technology (Bangor et al., 2008) • Normative data available	• Focuses on the whole system rather than specific aspects, making diagnostic judgments difficult
TAM	• Free to use • Generalizable as tested across different systems, user groups, and research settings (Davis, 1989)	N/A
UEQ	• Free to use • Covers a wide range of areas including pragmatic and hedonic aspects • Normative data available	• Quite time-consuming to complete
UMUX	• Free to use • Quick to complete • Normative data available	• Limited number of items • Focuses on the whole system rather than specific aspects, making specific diagnostic judgments difficult
UMUX-LITE	• Free to use • Very quick to complete • Normative data available	• Very limited number of items • Focuses on the whole system rather than specific aspects, making diagnostic judgments difficult

Questionnaire	Advantages	Disadvantages
USE	• Free to use • Can measure dimensions of usability in a variety of domains (e.g., software, hardware, services, and user support materials) to compare systems from different domains (Lund, 2001)	• Focuses on the whole system rather than specific aspects, making diagnostic judgments difficult • Quite time-consuming to complete

Note. N/A = Not applicable (No specific advantage or disadvantage has been identified).

Advantages and disadvantages include whether a license fee is required, coverage of items, generalizability, flexibility, availability of normative data, diagnostic ability versus just comparative, questionnaire design, specific questionnaire or subscales issues, completion time, and use with certain sample sizes. As the questionnaires are summarized in Table 2 alphabetically, rather than in order of strengths, each should be considered for any featured advantages or disadvantages, then these should be weighed up if comparing between questionnaires. Certain advantages may be considered beneficial to a study, for example availability of normative data, as it allows the researcher to interpret scores of a single system or product. Only some of the questionnaires above feature such norms, with some of these being publicly available and others being proprietary norms associated with a license.

Aspects of Usability Covered by Post-Study Questionnaires

Table 3 summarizes the aspects of usability covered by the questionnaires. The questionnaires are ordered first by the greatest number of usability categories (for emphasis), then alphabetically (for clarity).

Table 3. Summary of Aspects of Usability Covered by Post-Study Questionnaires

Questionnaire	Aspect of usability			
	Effectiveness	Efficiency	Satisfaction	Learnability
AttrakDiff 2	✓	✓	✓[a]	✓
CSUQ	✓	✓	✓	✓
meCUE	✓[a]	✓	✓[a]	✓
PSSUQ	✓	✓	✓	✓
QUIS	✓	✓	✓	✓[a]
SUMI	✓	✓	✓	✓[a]
TAM	✓[a]	✓	✓[a]	✓
UEQ	✓[a]	✓[a]	✓[a]	✓[a]
USE	✓[a]	✓[a]	✓[a]	✓[a]
EUCS	✓	✓	✓	
PUTQ	✓[a]	✓[a]		✓[a]
UMUX	✓	✓	✓	
SUS			✓[a]	✓[ab]
UMUX-LITE	✓		✓	

[a] Aspect of usability covered is either the whole scale or a subscale. [b] See discussion below for further details.

As presented in Table 3, nine questionnaires include all aspects of usability (AttrakDiff 2, CSUQ, meCUE, PSSUQ, QUIS, SUMI, TAM, UEQ, and USE), three were missing only one aspect (EUCS, PUTQ, and UMUX), with the SUS and UMUX-LITE only covering two aspects. This means that the majority of the questionnaires cover all, or nearly all, aspects but caution should be applied if using the latter two on their own. In addition, only some measure a specific aspect as an

individual subscale (e.g., a subscale only including efficiency questions), which may be important for a usability study. Furthermore, previous research on the SUS revealed Learnability to be a separate subscale (Lewis & Sauro, 2009), but this has since been challenged (Lewis & Sauro, 2017; Lewis et al., 2015).

Psychometric Quality of Post-Study Questionnaires

The following sections present the analysis for reliability, validity, and sensitivity.

Reliability Analysis

Table 4 summarizes the reliability scores for all relevant questionnaires, ordered first by the highest overall score (out of any studies), then the highest score in a subscale(s) (for further emphasis), and then alphabetically (for clarity). Also, within each questionnaire, the subscales have been ordered based on the highest value from any study, and each subscale's scores have been ordered with the highest first, highlighting relative strengths of individual subscales.

Table 4. Summary of Reliability Figures for Post-Study Questionnaires

Questionnaire	Reliability (*r*)	Sources
EUCS	Overall = .99, .92; Format = .98, .78; Accuracy = .91, .68; Content = .89, .83; Ease of use = .85, .52; Timeliness = .82, .78	Doll & Torkzadeh (1988); Doll et al. (1994)
TAM	Overall = .98, .95, .95; Usefulness = .98, .98, .95, .94, .93; Ease of use = .97, .95, .95, .94, .92	Davis (1989); Lah et al. (2020)
USE	Overall = .98; Ease of use = .95, .94; Usefulness = .93, .91; Satisfaction = .91, .88; Ease of learning = .90, .87	Gao et al. (2018)
CSUQ	Overall = .97, .97, .95; SysUse = .96, .95, .93; InfoQual = .93, .93, .91; IntQual = .91, .90, .89	Lewis (2002, 2018, 2019)
PSSUQ	Overall = .96; SysUse = .96; InfoQual = .92; IntQual = .83	Lewis (2002)
QUIS	Overall = .94	Chin et al. (1988)
SUS	Overall = .94, .94, .93, .92, .91, .91, .88; Usability = .91; Learnability = .70	Bangor et al. (2008); Lah et al. (2020); Lewis (2018, 2019); Lewis & Sauro (2009)
UMUX	94, .91, .89, .88, .87, .85, .81, .79	Finstad (2010); Lah et al. (2020); Lewis (2018, 2019); Lewis et al. (2013)
meCUE	Positive emotions = .94, .82; Negative emotions = .92, .88; Visual aesthetics = .91; Usability = .90, .89, .89; Social identity: Commitment = .86, .76; Social identity: Status = .84, .83; Usefulness = .83, .78	Minge et al. (2016); Minge et al. (2017)
SUMI	Global = .92; Affect = .85; Helpfulness = .83; Learnability = .82; Efficiency = .81; Control = .71	Kirakowski (1994)
UEQ	Attractiveness = .89, .86; Stimulation = .88, .76; Novelty = .84, .83; Perspicuity = .82, .71; Efficiency = .79, .73; Dependability = .69, .65	Laugwitz et al. (2008)
UMUX-LITE	.86, .84, .83, .82, .79, .76, .73, .69	Lah et al. (2020); Lewis (2018, 2019); Lewis et al. (2013); Lewis et al. (2015)
PUTQ	.81-.59	Lin et al. (1997)
AttrakDiff 2	ATT[a] = .70; HQS = .95, 90, .76, .55; PQ = .91, .86, .85, .85, .83; HQI = .86, .83, .73, .45	Hasenzahl (2004); Hassenzahl et al. (2003); Hassenzahl & Sandweg (2004); Isleifsdottir & Larusdottir (2008)

Note. Unless mentioned, reliability statistics are for the overall scale, not specific subscales.
[a] While the AttrakDiff 2 does not have a traditional overall score (e.g., averaged from subscales), the ATT subscale measures the overall attraction of a system, so it is considered an overall score here.

The EUCS achieved the highest overall reliability, followed by the TAM, then the USE, with the CSUQ and PSSUQ close behind. All scales achieved a suitable level of reliability (at least with the highest scores found), except there is some concern regarding the EUCS, PUTQ, AttrakDiff 2, UEQ, and UMUX-LITE. While the EUCS achieved the highest overall reliability, the Accuracy and Ease of use subscales' lowest identified scores fell below the suggested amount of .70. This might be explained by the different method used to estimate reliability, as another study identified suitable reliability scores for these subscales. At the opposite end, the AttrakDiff 2 had the lowest overall score, just reaching the cut-off of .70. Also, one of the findings for each of the HQI and HQS subscales was especially low. However, this might be

overshadowed by numerous other findings for these subscales being much higher. For the PUTQ, its range of scores had the second lowest (but still suitable) maximum reliability compared to other questionnaires, with its minimum score falling below .70. While the UMUX-LITE is next lowest, and its lowest score just falling below the cut-off, its overall reliability is noted as being excellent for a two-item measure (Lewis et al., 2015). However, it should be noted that some questionnaires (e.g., the EUCS, SUS, meCUE, AttrakDiff 2, and UEQ) had a greater discrepancy between the highest and lowest values between or within subscales than other questionnaires.

Validity Analysis

Content validity does not involve correlations, and the absolute values of criterion validity are not as important as for reliability. For these reasons, the different types of validity may be of more importance. Table 5 summarizes the validity categories for each relevant questionnaire, ordered first on the number of types, then on the number of broad sources of criterion validity (e.g., another questionnaire or other measure), and then alphabetically. For the USE, self-predicted and actual usage correlations were reported to be strong, but no figures were provided (Lund, 2001). The SUS factor analysis revealed it included two factors (Lewis & Sauro, 2009), but doubt has been raised on its dimensionality (Lewis et al., 2015). A later confirmatory factor analysis (Lewis & Sauro, 2017) suggested that the Learnability subscale is likely an artifact, with an apparent two-factor model being due to a mix of positively and negatively-toned items. For the SUS criterion validity with the CUSI Affect subscale and the QUIS, exact figures are unknown, as only the range of correlations between all three scales was referenced. Also, regarding task performance, in Peres et al.'s (2013) study, seven out of eight studies were non-significant correlations (all studies combined reduced the relationship, $r = .22$). In addition to correlations between the SUS, UMUX, UMUX-LITE, and the CSUQ, Lewis (2018) found evidence that when measures were converted to match the SUS' scale, there was a close correspondence between the measures, further strengthening criterion validity. For the QUIS' criterion validity, as mentioned above, only the range of correlations is known for this and the SUS. In addition to the UEQ correlations found with completion time, Laugwitz et al. (2008) also found no correlation for attractiveness, stimulation, and novelty, as expected (i.e., discriminant validity), which further provided criterion validity. Further to the expected correlations they found with the AttrakDiff 2, they also found a relationship between UEQ Dependability and AttrakDiff 2 HQI, but as this was not expected, it was not included as criterion validity. Further to the expected meCUE correlations with number of completed tasks, Minge et al. (2017) also found no correlation with the Visual aesthetics, Status, Commitment, Positive emotions, and Negative emotions subscales, also as expected, thus providing additional criterion validity.

Table 5. Summary of Validity Categories for Post-Study Questionnaires

Questionnaire	Type of validity		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
meCUE	✓ (Minge et al., 2017)	✓ (Minge et al., 2016; Minge et al., 2017)	✓ **AttrakDiff 2 PQ:** Usefulness (r = .64), Usability (r = .90, .87), Visual aesthetics (r = .57), Status (r = .46), Commitment (r = .53); **HQI:** Usefulness (r = .62), Usability (r = .52), Visual aesthetics (r = .67), Status (r = .51), Commitment (r = .58); **HQS:** Usefulness (r = .40), Usability (r = .37), Visual aesthetics (r = .72), Status (r = .51), Commitment (r = .50); **ATT:** Usefulness (r = .67), Usability (r = .68), Visual aesthetics (r = .77), Status (r = .55), Commitment (r = .64), Global attractiveness (r = .56) (Minge et al., 2017)
			UEQ Efficiency: Usefulness (r = .61), Usability (r = .90, .65), Visual aesthetics (r = .55), Status (r = .35), Commitment (r = .44); **Perspicuity:** Usefulness (r = .62), Usability (r = .86, .85), Visual aesthetics (r = .48), Status (r = .37), Commitment (r = .44); **Dependability:** Usefulness (r = .69), Usability (r = .78, .73), Visual aesthetics (r = .54), Status (r = .43), Commitment (r = .54); **Stimulation:** Usefulness (r = .62), Usability (r = .61), Visual aesthetics (r = .72), Status (r = .54), Commitment (r = .58); **Novelty:** Usefulness (r = .36), Usability (r = .40), Visual aesthetics (r = .67), Status (r = .48), Commitment (r = .45); **Attractiveness:** Usefulness (r = .68), Usability (r = .70), Visual aesthetics (r = .74), Status (r = .54), Commitment (r = .60), Global attractiveness (r = .89) (Minge et al., 2017)
			Visual aesthetics / Classical aesthetics: Usefulness (r = .46), Usability (r = .52), Visual aesthetics (r = .70), Status (r = .42), Commitment (r = .43); **Expressive aesthetics:** Usefulness (r = .43), Usability (r = .40), Visual aesthetics (r = .75), Status (r = .56), Commitment (r = .51) (Minge et al., 2017)
			Number of completed tasks: Usefulness (r = .32), Usability (r = .34) (Minge et al., 2017)
			PANAS Positive affect: Positive emotions (r = .51, .47), Negative emotions (r = -.39); **Negative affect:** Positive emotions (r = -.26), Negative emotions (r = .72, .63) (Minge et al., 2017)
			PANAS Positive affect: Positive emotions (r = .51), Negative emotions (r = -.39); **Negative affect:** Positive emotions (r = -.26), Negative emotions (r = .63) (Minge et al., 2017)
			SAM Arousal: Positive emotions (r = -.22), Negative emotions (r = .35); **Valence:** Positive emotions (r = .66), Negative emotions (r = .63) (Minge et al., 2017)

Questionnaire	Type of validity		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
TAM	✓ (Davis, 1989)	✓ (Davis, 1989; Lah et al., 2020)	✓ **SUS:** Overall (*r* = .90, .80, 70), Usefulness (*r* = .83, .61, .52), and Ease of use (*r* = .90, .84, .78) (Lah et al., 2020) **UMUX:** Overall (*r* = .90, .76, .63), Usefulness (*r* = .86, .59, .45), and Ease of use (*r* = .87, .78, .71) (Lah et al., 2020) **UMUX-LITE:** Overall (*r* = .89, .77, .67), Usefulness (*r* = .85, .62, .49), and Ease of use (*r* = .87, .78, .75) (Lah et al., 2020) **Actual usage:** Usefulness (*r* = .68, .56) and Ease of use (*r* = .48, .32) (Davis, 1989) **Self-predicted usage:** Ease of use (*r* = .47) and Usefulness (*r* = .71, .59) (Davis, 1989)
AttrakDiff 2	✓ (Hassenzahl et al., 2003)	✓ (Hassenzahl et al., 2003)	✓ **Item measuring Beauty:** HCI (*r* = .61) (Hassenzahl, 2004) **Item measuring Goodness:** HQI (*r* = .49) and PQ (*r* = .41) (Hassenzahl, 2004) **SMEQ:** PQ (Hassenzahl & Sandweg, 2004) **meCUE Usefulness:** PQ (*r* = .64), HQI (*r* = .62), HQS (*r* = .40), ATT (*r* = .67); **Usability:** PQ (*r* = .90, .87), HQI (*r* = .52), HQS (*r* = .37), ATT (*r* = .68); **Visual aesthetics:** PQ (*r* = .57), HQI (*r* = .67), HQS (*r* = .72), ATT (*r* = .77); **Status:** PQ (*r* = .46), HQI (*r* = .51), HQS (*r* = .51), ATT (*r* = .55); **Commitment:** PQ (*r* = .53), HQI (*r* = .58), HQS (*r* = .50), ATT (*r* = .64); **Global attractiveness:** ATT (*r* = .56) (Minge et al., 2016)
UEQ	✓ (Laugwitz et al., 2008)	✓ (Laugwitz et al., 2008)	✓ **AttrakDiff 2 PQ:** Perspicuity (*r* = .73), Efficiency (*r* = .59), Dependability (*r* = .54); **HQS:** Stimulation (*r* = .72), Novelty (*r* = .64) (Laugwitz et al., 2008[a]) **meCUE Usefulness:** Efficiency (*r* = .61), Perspicuity (*r* = .62), Dependability (*r* = .69), Stimulation (*r* = .62), Novelty (*r* = .36), Attractiveness (*r* = .68); **Usability:** Efficiency (*r* = .90, .65), Perspicuity (*r* = .86, .85), Dependability (*r* = .78, .73), Stimulation (*r* = .61), Novelty (*r* = .40), Attractiveness (*r* = .70); **Visual aesthetics:** Efficiency (*r* = .55), Perspicuity (*r* = .48), Dependability (*r* = .54), Stimulation (*r* = .72), Novelty (*r* = .67), Attractiveness (*r* = .74); **Status:** Efficiency (*r* = .35), Perspicuity (*r* = .37), Dependability (*r* = .43), Stimulation (*r* = .54), Novelty (*r* = .48), Attractiveness (*r* = .54); **Commitment:** Efficiency (*r* = .44), Perspicuity (*r* = .44), Dependability (*r* = .54), Stimulation (*r* = .58), Novelty (*r* = .45), Attractiveness (*r* = .60); **Global attractiveness:** Attractiveness (*r* = .89) (Minge et al., 2016)

Questionnaire	Type of validity		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
			Task completion time: Perspicuity (*r* = -.66), Efficiency (*r* = -.73), Dependability (*r* = -.65) (Laugwitz et al., 2008[a])
USE	✓ (Lund, 2001)	✓ (Gao et al., 2018)	✓ **SUS:** Usefulness (*r* = .69, .60), Ease of use (*r* = .81, .78), Ease of learning (*r* = .78, .71), and Satisfaction (*r* = .71, .66) (Gao et al., 2018) **Self-predicted usage:** Satisfaction (Lund, 2001[a]) **Actual usage:** Satisfaction (Lund, 2001[a])
PSSUQ	✓ (Lewis, 2002)	✓ (Lewis, 1995, 2002)	✓ **ASQ** (*r* = .80, Lewis, 1995) **Completion rates** (*r* = .40, Lewis, 1995)
CSUQ	✓ (Lewis, 1995)	✓ (Lewis, 1995, 2018, 2019)	✓ **SUS:** Overall (*r* = .87, Lewis, 2019; *r* = .76, Lewis, 2018), SysUse (*r* = .74), InfoQual (*r* = .65), IntQual (*r* = .68) (Lewis, 2018)
SUS		✓** (Bangor et al., 2008; Lah et al., 2020; Lewis, 2018, 2019; Lewis & Sauro, 2009, 2017)	✓ **SUMI** (*r* = .79, Sauro, 2011a) **WAMMI** (*r* = .95, Sauro, 2011a) **UMUX** (*r* = .96, Lewis et al., 2013; *r* = .92, Lah et al., 2020; *r* = .90, Finstad, 2010; *r* = .90, Lewis, 2019; *r* = .86, Lah et al., 2020; *r* = .79, Lewis, 2018; *r* = .78, Lah et al., 2020; *r* = .72, .55, Borsci et al., 2015) **UMUX-LITE** (*r* = .89, Lah et al., 2020; *r* = .86, Lewis, 2019; *r* = .85 [positive version of SUS], Lewis et al., 2013; *r* = .83, Lewis et al., 2015; *r* = .82, Lah et al., 2020; *r* = .81, Lewis et al., 2013; *r* = .74, Lah et al., 2020; *r* = .74, Lewis, 2018; *r* = .66, .45, Borsci et al., 2015) **CUSI Affect** (*r* = .67-.74, Wong & Rengger, 1990, as cited by Kirakowski, 1994[a]); **Competence** (*r* = .58, Wong & Rengger, 1990, as cited by Kirakowski, 1994) **QUIS** (*r* = .67-.74, Wong & Rengger, 1990, as cited by Kirakowski, 1994[a]) **SEQ** (*r* = -.57, Sauro & Dumas, 2009) **SMEQ** (*r* = -.60, Sauro & Dumas, 2009) **UME** (*r* = -.32, Sauro & Dumas, 2009) **SUPR-Q Usability** (*r* = .96, Sauro & Lewis, 2012) **USE Usefulness** (*r* = .69, .60); **Ease of use** (*r* = .81, .78); **Ease of learning** (*r* = .78, .71); **Satisfaction** (*r* = .71, .66) (Gao et al., 2018) **CSUQ Overall** (*r* = .90, Lewis, 2019; *r* = .76, Lewis, 2018); **SysUse** (*r* = .74); **InfoQual** (*r* = .65); **IntQual** (*r* = .68) (Lewis, 2018)

Questionnaire	Type of validity		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
			TAM Overall (*r* = .90, .80, .70); **Usefulness** (*r* = .83, .61, .52); **Ease of use** (*r* = .90, .84, .78) (Lah et al., 2020) **Ratings of systems** (*r* = .82, Bangor et al., 2009; *r* = .81, Bangor et al., 2008) **Task performance** (*r* = .63, Peres et al., 2013[a]; *r* = .24, Sauro, 2011b). **Overall Experience** (*r* = .89, .80, Lah et al., 2020; *r* = .67, Lewis et al., 2015; *r* = .64, Lah et al., 2020) **LTR** (*r* = .88, .75, Lah et al., 2020; *r* = .71, Lewis et al., 2015; *r* = .60, Lah et al., 2020)
UMUX		✓ (Finstad, 2010; Lah et al., 2020; Lewis, 2018, 2019; Lewis et al., 2013, 2015)	✓ **SUS** (*r* = .96, Lewis et al., 2013; *r* = .92, Lah et al., 2020; *r* = .90, Finstad, 2010; *r* = .90, Lewis, 2019; *r* = .86, Lah et al., 2020; *r* = .79, Lewis, 2018; *r* = .78, Lah et al., 2020; *r* = .72, .55, Borsci et al., 2015) **TAM Overall** (*r* = .90, .76, .63); **Usefulness** (*r* = .86, .59, .45); **Ease of use** (*r* = .87, .78, .71) (Lah et al., 2020) **Overall Experience** (*r* = .92, .79, .65) (Lah et al., 2020) **LTR** (*r* = .89, .73, .61) (Lah et al., 2020)
UMUX-LITE		✓ (Lah et al., 2020; Lewis, 2018, 2019; Lewis et al., 2013)	✓ **SUS** (*r* = .89, Lah et al., 2020; *r* = .86, Lewis, 2019; *r* = .85 [positive version of SUS], Lewis et al., 2013; *r* = .83, Lewis et al., 2015; *r* = .82, Lah et al., 2020; *r* = .81, Lewis et al., 2013; *r* = .74, Lah et al., 2020; *r* = .74, Lewis, 2018; *r* = .66, .45, Borsci et al., 2015) **TAM Overall** (*r* = .89, .76, .67); **Usefulness** (*r* = .85, .59, .49); **Ease of use** (*r* = .87, .78, .75) (Lah et al., 2020) **Overall Experience** (*r* = .90, .78, Lah et al., 2020; *r* = .72, Lewis et al., 2015; *r* = .66, Lah et al., 2020) **LTR** (*r* = .87, .75, Lah et al., 2020; *r* = .74, .73, Lewis et al., 2013; *r* = .72, Lewis et al., 2015; *r* = .62, Lah et al., 2020)
QUIS		✓ (Chin et al., 1988)	✓ **CUSI Affect** (*r* = .67-.74, Wong & Rengger, 1990, as cited by Kirakowski, 1994[a]); **Competence** (*r* = .38, Wong & Rengger, 1990, as cited by Kirakowski, 1994) **SUS** (*r* = .67-.74) (Wong & Rengger, 1990, as cited by Kirakowski, 1994[a]) **PUTQ** (*r* = .87, Lin et al., 1997)
EUCS		✓ (Doll et al., 1994)	✓ **Item measuring system satisfaction:** Overall (*r* = .76), Content (*r* = .69), Accuracy (*r* = .55), Format (*r* = .60), Ease of use (*r* = .58), and Timeliness (*r* = .60) (Doll & Torkzadeh, 1988)

Questionnaire	Type of validity		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
PUTQ	✓ (Lin et al., 1997)		✓ **QUIS** (*r* = .87, Lin et al., 1997)
SUMI	✓ (Kirakowski, 1994)	✓ (Kirakowski, 1994)	

Note. Unless mentioned, criterion validity is for the overall scale, not specific subscales. PANAS = Positive and Negative Affect Schedule. SAM = Self-Assessment Manikin. WAMMI = Website Analysis and MeasureMent Inventory. CUSI = Computer User Satisfaction Inventory. SUPR-Q = Standardized User Experience Percentile Rank Questionnaire. Overall experience = 5-point semantic differential scale (1 = *I hate it* to 5 = *I love it*). LTR = Likelihood-to-Recommend.
[a] There is extra discussion or validity was calculated in a specific way (see above the table for detail).

The meCUE, TAM, AttrakDiff 2, UEQ, USE, PSSUQ, and CSUQ covered all three categories of validity, with all except the CSUQ having multiple types of criterion validity. The remainder covered two categories of validity. However, just considering criterion validity, the SUS has a clear advantage with 17 different variables. Out of the rest, the EUCS, PUTQ, and SUMI were weakest with just one or no example of criterion validity.

Sensitivity Analysis

Table 6 summarizes the sensitivity categories for each relevant questionnaire, ordered first by the number of categories, then alphabetically. While this review categorizes sensitivity broadly, Lewis (2002) conducted numerous specific tests, including a study where PSSUQ data was taken from the developer of the product, the stage of development, the type of product, and the type of evaluation. This means that multiple types of sensitivity fall into individual categories. While Chin et al. (1988) found the QUIS to be sensitive to system differences, Lin et al. (1997) found it did not differentiate between two systems. The PUTQ was only sensitive to system differences at $p < .10$, however, Lin et al. referred to previous evidence and their own study, which supported the specified difference between the systems. In another study, the PUTQ was insensitive to expertise but not to gender differences (De La Cruz, n.d.); however, the latter is not a relevant category focused on here. Lund (2001) found the USE to be insensitive to product differences for the usefulness subscale. The AttrakDiff 2 was sensitive to system differences for PQ and HQS but not HQI (Hassenzahl, 2004), and Hassenzahl et al. (2003) found some evidence of such sensitivity (an interaction between website and subscale). However, Hassenzahl (2004) did find it was sensitive for all subscales for overall ratings of "beautiful" and "ugly." In addition, while it was sensitive to pre- and post-use differences for HQS and ATT, it was not for HQI and PQ (Isleifsdottir & Larusdottir, 2008), and Hassenzahl found no difference. For the UEQ, Schrepp et al. (2014) found system differences for Attractiveness, Perspicuity, Efficiency, Dependability, but not Stimulation or Novelty.

Table 6. Summary of Sensitivity Categories for Post-Study Questionnaires

Questionnaire	Type of sensitivity				
	System differences	**Frequency of use**	**Expertise**	**Task/ Evaluation differences**	**Overall ratings**
AttrakDiff 2	✓ (Hassenzahl, 2004[a]; Hassenzahl et al., 2003)	✓ (Hassenzahl, 2004[a]; Isleifsdottir & Larusdottir, 2008[a])			✓ (Hassenzahl, 2004[a])
PSSUQ	✓ (Lewis, 2002[a])		✓ (Lewis, 1995)	✓ (Lewis, 2002[a])	
SUS	✓ (Bangor et al., 2008; Gao et al., 2018; Lewis & Sauro, 2009)	✓ (Borsci et al., 2015; Kortum & Johnson, 2013; Lewis et al., 2015; McLellan et al., 2012; Sauro, 2011c)	✓ (Lewis et al., 2015)		
CSUQ	✓ (Tullis & Stetson, 2004)		✓ (Lewis, 1995)		
QUIS	✓ (Chin et al., 1988; Lin et al., 1997[a])				✓ (Chin et al., 1988)
SUMI	✓ (Kirakowski, 1994)		✓ (Kirakowski, 1994)		
UMUX	✓ (Finstad, 2010)	✓ (Borsci et al., 2015)			
UMUX-LITE		✓ (Borsci et al., 2015; Lewis et al., 2015)	✓ (Lewis et al., 2015)		
meCUE	✓ (Minge et al., 2017)				
PUTQ	✓ (Lin et al., 1997[a])		(De La Cruz, n.d.[a])		
TAM	✓ (Davis, 1989)				
UEQ	✓ (Schrepp et al., 2014[a])				
USE	✓ (Lund, 2001[a])				
EUCS					

Note. Unless mentioned, sensitivity categories are for the overall scale, not specific subscales.
[a] There is extra discussion, or sensitivity was calculated in a specific way (see above table for detail).

The AttrakDiff 2, PSSUQ, and SUS feature the greatest number (three) of types of sensitivity, supporting a previous finding that the most sensitive post-study measure is the SUS, then the PSSUQ (Sauro & Lewis, 2012). Five questionnaires covered two types, with five more covering one. For one of the latter, the PUTQ had some potential issues with system differences not achieving significance at the usual level of $p < .05$, and it was insensitive to expertise differences yet found unexpected gender differences. In addition, the UEQ and USE were not sensitive for all subscales. However, the EUCS was weakest overall with no sensitivity evidence found.

Post-Task Questionnaires Analysis

This section includes those questionnaires that were identified as being a post-task questionnaire type.

Structure and Content of Post-Task Questionnaires

Table 7 summarizes the key post-task questionnaires used for assessing usability, ordered alphabetically by questionnaire name, as at this stage no questionnaire takes precedence over another.

Table 7. Summary of Post-Task Questionnaires with General Details

Questionnaire (acronym)	Questionnaire general details			
	Subscales	**Type of scale**	**Number of points**	**Number of items**
After Scenario Questionnaire (ASQ)	N/A	LS	7	3
Expectation Ratings (ER)	N/A	SDS	7/5	2
Single Ease Question (SEQ)	N/A	SDS	7/5	1
Subject Mental Effort Question (SMEQ)	N/A	SDS	150	1
Usability Magnitude Estimation (UME)	N/A	SDS	100	1

Note. N/A = Not applicable (The questionnaire has no subscales; it just includes an overall score). LS = Likert scale. SDS = Semantic differential scale.

The ASQ (available in Lewis, 1991) was developed with items measuring the ease and quickness of completing the scenario, along with support from information on how to achieve the task (Kirakowski, 1994). The SEQ is similar to ASQ Item 1, basically asking users to rate how easy overall it was to complete the task (available in Sauro & Lewis, 2012). The SMEQ (available in Sauro & Lewis, 2012) is another single item scale, ranging from 0 to 150 in addition to having nine descriptive labels with "Not at all hard to do" existing just above 0 and "Tremendously hard to do" existing just above 110. Participants draw a line through the scale (paper version) or move a slider (online version) to represent their mental effort in task completion.

The ER (available in Albert & Dixon, 2003) involves expectation ratings, the difference between how easy a task was experienced to be and perceived to be beforehand. Ratings are similar to the SEQ, except participants rate before and after the task, making this a two-item measure (Sauro & Lewis, 2012). The UME relates to magnitude estimation, judgment of how intense a stimulus is compared to a baseline stimulus (e.g., the ratio of the brightness of a stimulus light to a reference light). The UME (Cordes, 1984; McGee, 2003; Sauro & Dumas, 2009) measures the ratio of a task/product's difficulty to another (i.e., a perceived difficulty of 100 is twice as difficult as one with a score of 50; available in Sauro & Lewis, 2012). Normally, participants receive pre-training in magnitude estimation, such as comparing line length to a reference line (McGee, 2003). For assessing usability, a very easy baseline task is compared with the main task (Cordes, 1984), with scores converted to a consistent ratio scale line for comparison (McGee, 2003).

Advantages and Disadvantages of Post-Task Questionnaires

Table 8 summarizes the advantages and disadvantages of each relevant post-task questionnaire. Questionnaires are ordered alphabetically as while some have more advantages and less disadvantages than others, the relevance or importance of each is subject to the individual usability study.

Table 8. Summary of Advantages and Disadvantages of Post-Task Questionnaires

Questionnaire	Advantages	Disadvantages
ASQ	• Free to use • Very quick to complete	• Limited number of items • Limited generalizability (see Lewis, 1995)
ER	• Free to use • Before-and-after rating allows for plotting scores on a graph to map four different scenarios (Albert & Tullis, 2013) includes 1) Task was expected to be difficult but was not experienced as such (opportunity to promote such features) 2) Task was both predicted and experienced as difficult (opportunity to improve) 3) Both tasks considered as easy (consider leaving alone) 4) Tasks were difficult but expected to be easy (opportunity for focused improvement to remove dissatisfaction)	N/A
SEQ	• Free to use • Very quick to complete • Functions well despite its simplicity, and it performs well in comparison to more complicated measures, e.g., the SMEQ and UME (Sauro, 2012) • Normative data available	• Very limited with only one item
SMEQ	• Free to use • Very quick to complete • The placement of labels on the scale is based on psychometric calibration with tasks (Sauro & Dumas, 2009)	• Very limited with only one item
UME	• Free to use • Continuous as it has no upper bound limit (Sauro & Dumas, 2009), which is believed to overcome limitations of other forms of usability measurement, such as with scale items having fixed endpoints that could restrict responses (Sauro & Lewis, 2012)	• Evidence that the UME's full potential was not being used by participants—displayed a limited range of responses (Sauro & Dumas, 2009) • Confusion that participants appear to have regarding making ratio judgments (Sauro & Dumas, 2009; Tedesco & Tullis, 2006)

Note. N/A = No specific advantage or disadvantage has been identified.

Advantages and disadvantages include whether a license fee is required, coverage of items, generalizability, availability of normative data, specific questionnaire design, and completion time. As per the post-study questionnaires, such advantages and disadvantages of each questionnaire should be considered and weighed up when comparing specific questionnaires.

Aspects of Usability Covered by Post-Task Questionnaires

Table 9 presents a summary of aspects of usability. The questionnaires are ordered first by the greatest number of usability categories (for emphasis), then alphabetically (for clarity).

Table 9. Summary of Aspects of Usability Covered by Post-Task Questionnaires

Questionnaire	Aspect of usability			
	Effectiveness	Efficiency	Satisfaction	Learnability
ASQ		✓		✓
ER		✓[a]		
SEQ		✓[a]		
SMEQ		✓[a]		
UME		✓[a]		

[a] Aspect of usability covered is the whole scale.

As evident in Table 9, the ASQ covers the most categories, with two aspects covered (but not measured as separate subscales). The remaining measures are equal in coverage; however, they only cover one aspect, efficiency, so a usability study would require a further measure(s) to have a broader assessment of usability.

Psychometric Quality of Post-Task Questionnaires

The following sections present the analysis for reliability, validity, and sensitivity.

Reliability Analysis

Table 10 summarizes the reliability scores for all relevant questionnaires, ordered first by the highest score (out of any studies), then further scores identified from studies (for further emphasis), and then alphabetically (for clarity). For the ER, reliability cannot be determined from correlations between the two ratings, as expectations and experiences could differ. It is unknown if test-retest reliability has been conducted, or is suitable, as the participant's expectations would need to be novel. Tedesco and Tullis (2006) explored reliability by taking 1,000 random samples (sizes from 3–29) from their dataset then correlating sub-samples with the total sample score across multiple tasks. The quoted figure is for samples of 23 and over. For the SEQ, UME, and ASQ (using the average of items 1 and 2), Tedesco and Tullis explored reliability in the same way as described for the ER. The creators of the SMEQ claim it is reliable (Sauro & Dumas, 2009), through comparing scores between numerous conditions providing test-retest reliability (Zijlstra, 1993).

Table 10. Summary of Reliability Figures for Post-Task Questionnaires

Questionnaire	Reliability (*r*)	Sources
ER	.95[a]	Tedesco and Tullis (2006)
SEQ	.95 [a]	Tedesco and Tullis (2006)
UME	.95[a]	Tedesco and Tullis (2006)
ASQ	.95[a], .90	Tedesco and Tullis (2006), Lewis (1995)
SMEQ	.88, .81, .71, .58	Zijlstra (1993)

[a] Reliability was calculated in a specific way (see above table for details).

Aside from the SMEQ, all questionnaires achieved high reliability with the highest score equivalent, and only the ASQ having a somewhat lower score as identified elsewhere. Thus, out of these, there is no preference with regards to an individual post-task measure. The SMEQ had

a lower range of scores with one falling below the suggested minimum amount of .70. However, the authors noted that lower scores should be expected, due to the natural variability in mental effort (Zijlstra, 1993).

Validity Analysis

Table 11 presents a summary of the validity categories for each relevant questionnaire, ordered first on the number of types, then on the number of sources of criterion validity, and then alphabetically. The ASQ (average of items 1 and 2) had a high correlation with errors made and the UME, but in both cases the authors noted limited statistical power from the small sample (see Sauro & Dumas, 2009). For Tedesco and Tullis' (2006) correlation (also average of items 1 and 2) with performance efficiency (combination of completion rates and completion time), only the range of correlations between numerous usability measures and performance efficiency was referenced. For the UME, in addition to the ASQ correlation, Sauro and Dumas (2009) smaller study's task completion rates, time, and errors (the only significant correlation, $r = .78$) correlations were contradictory to their larger study (figures quoted in Table 11), possibly due to the limited power. For the SEQ, the correlation is a range as features in the Tedesco and Tullis' study. Sauro and Dumas' correlation with errors was only significant at $p < .10$. Similarly, Sauro and Dumas only found the SMEQ correlation with completion rates significant at $p < .10$. Kirakowski and Cierlik (1998) compared two websites and found the one requiring less mental effort corresponded to greater efficiency (completion time divided by an expert's completion time). Tedesco and Tullis' ER correlation with performance efficiency (see above) only involved the post-task question.

Table 11. Summary of Validity Categories for Post-Task Questionnaires

Questionnaire	**Type of validity**		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
ASQ	✓ (Lewis, 1995)	✓ (Lewis, 1995)	✓ **UME** ($r = .84$, Sauro & Dumas, 2009[a]) **Task performance** ($r = .46$-$.37$, Tedesco & Tullis, 2006[a]; $r = .40$, Lewis, 1995), **errors made** ($r = .73$, Sauro & Dumas, 2009[a])
UME			✓ **SMEQ** ($r = .85$, Sauro & Dumas, 2009) **SEQ** ($r = .96$, Sauro & Dumas, 2009) **SUS** ($r = .32$, Sauro & Dumas, 2009) **ASQ** ($r = .84$, Sauro & Dumas, 2009[a]) **Task completion time** ($r = -.91$, Sauro & Dumas, 2009[a]; $r = -.24$, McGee, 2003), **completion rates** ($r = -.05$, Sauro & Dumas, 2009[a]), **number of clicks** ($r = -.39$, McGee, 2003), **errors** ($r = .78$, $-.24$, Sauro & Dumas, 2009[a]; $r = -.20$, McGee, 2003), and **assists** ($r = -.19$, McGee, 2003)
SEQ			✓ **SUS** ($r = -.57$, Sauro & Dumas, 2009) **SMEQ** ($r = .94$, Sauro & Dumas, 2009) **UME** ($r = .96$, Sauro & Dumas, 2009) **Performance efficiency** ($r = .46$-$.37$, Tedesco & Tullis, 2006[a]), **completion times** ($r = -.90$), **errors made** ($r = -.84$), and **completion rates** ($r = .22$, Sauro & Dumas, 2009[a])

Questionnaire	Type of validity		
	Content validity	**Factorial validity**	**Criterion validity (Scale/measure correlated with)**
SMEQ			✓ **SUS** (*r* = −.60, Sauro & Dumas, 2009) **SEQ** (*r* = .94, Sauro & Dumas, 2009) **UME** (*r* = .85, Sauro & Dumas, 2009) **Completion rates** (*r* = .88), **completion time** (*r* = −.82, **errors made** (*r* = −.72) (Sauro & Dumas, 2009 [a]), and **efficiency** (Kirakowski & Cierlik, 1998[a])
ER			✓ **Performance efficiency** (*r* = .46, Tedesco & Tullis, 2006[a])

Note. Unless mentioned, criterion validity is for the overall scale, not specific subscales.
[a] There is extra discussion, or validity was calculated in a specific way (see above table for detail).

The ASQ covered all three categories of validity and more than one type of criterion validity. All other questionnaires just included criterion validity, with the UME having a slight advantage over the SEQ and SMEQ (5 compared to 4 examples). The ER was weakest overall, with just one aspect of criterion validity.

Sensitivity Analysis

Table 12 presents a summary of the sensitivity categories for each relevant questionnaire, ordered first by the number of categories, then alphabetically. The ASQ was found to be insensitive to level of expertise (Lewis, 1995), and Sauro and Dumas (2009) found the UME to be insensitive to differences of system experience. While Kirakowski and Cierlik (1998) noted the SMEQ scored differently, it is unclear if it is significantly sensitive to system differences.

Table 12. Summary of the Sensitivity Categories for Post-Task Questionnaires

Questionnaire	Type of sensitivity				
	System differences	**Frequency of use**	**Expertise**	**Task/Evaluation differences**	**Overall ratings**
ASQ	✓ (Lewis, 1995)		(Lewis, 1995[a])	✓ (Lewis, 1995; Tedesco & Tullis, 2006)	
UME	✓ (Cordes, 1984; Sauro & Dumas, 2009)	(Sauro & Dumas, 2009[a])		✓ (Sauro & Dumas, 2009)	
ER				✓ (Tullis & Stetson, 2004)	
SEQ				✓ (Tedesco & Tullis, 2006)	
SMEQ	✓ (Kirakowski & Cierlik, 1998[a])				

Note. [a] There is extra discussion, or sensitivity was calculated in a specific way (see above for detail).

The ASQ and UME have the strongest evidence of sensitivity, each featuring two types. However, both were found to be insensitive to an additional category of expertise and experience, respectively. All others featured one type of sensitivity, with the SMEQ weakest as there is doubt on whether this was statistically significant.

Discussion

Sauro and Lewis (2012) broadly categorized questionnaires as post-study, post-task, website, and other. Questionnaires have been categorized here in a similar way but only using the first two categories as it was decided not to include those measuring websites, and some previously categorized as "other" were considered as being relevant for the post-study category. The remainder did not fit the inclusion criteria and were thus rejected. The most relevant category should be based on the study context to help guide further decisions as to the most appropriate measure(s) to use. Once decided, researchers need to consider the questionnaire aspects most relevant to their study and system, then consider which questionnaires are strongest in assessing usability. Questionnaire considerations should include general structure and content details (including the wording of individual items), specific advantages or disadvantages, aspects of usability covered, and psychometric support.

Individual advantages and disadvantages should be considered when deciding the use of a specific questionnaire or when comparing questionnaires. Some have specific constraints, such as flexibility and involving a license fee, but may come with associated advantages. For example, the SUMI is a commercial questionnaire service that compares scores with normative databases (Sauro & Lewis, 2012). However, some normative data have been reported for non-commercial scales, such as the SUS (Bangor et al., 2008; Sauro, 2011a), PSSUQ and CSUQ (Lewis, 2002). Additionally, correspondence of scores found between the UMUX, UMUX-LITE, CSUQ (and thus also the PSSUQ) with the SUS (Lewis, 2018, 2019) allows for such measures to coincide with, and potentially benefit from, norms identified for the SUS. The availability of normative data is important as Sauro and Lewis (2012) highlighted that standardized questionnaires' scores do not mean anything inherently, instead they are useful in comparing between systems or study conditions. Despite this, studies have attempted to analyze figures to aid diagnostic interpretation, such as with the SUS (Bangor et al., 2008, 2009). The questionnaire content will determine its ability at diagnosing specific issues with a system, meaning some questionnaires will be more successful than others.

As Frøkjær et al. (2000) noted, it is important to consider all usability aspects when designing a usability study, thus all these aspects are also important when it comes to questionnaire choice. This ultimately depends on the researcher's need for the inclusion of subjective aspects of usability, as others may be measured objectively (e.g., assessing task efficiency). As detailed in the current review, all aspects were not represented by all questionnaires, meaning each should be considered based on the needs of the researcher.

Lastly, for psychometric support, Sauro and Lewis (2012) noted that reliability, validity, and sensitivity had been established for all the questionnaires they included, thus having potential value for usability evaluations. Similar findings are presented here. However, as made apparent from ranking questionnaires based on the identified evidence, there is a difference between some questionnaires and only some types of validity and sensitivity were represented.

The limitations of this review include the questionnaire selection. The criteria chosen were decided upon in order to include the greatest number of questionnaires applicable across a broad range of systems, rather than being designed specifically for one system. This means some measures may not be included that cover some usability aspects, and those included differ slightly from those looked at by Sauro and Lewis (2012). A systematic review may also produce a somewhat different range of questionnaires to consider. Lastly, while the adopted approach enables a holistic review of questionnaires across a range of aspects, it does not provide a quick answer for the best questionnaire(s) in a specific context. The strengths of a questionnaire would need to be considered in relation to the system, usability study design, and overall context.

For overall post-study questionnaire recommendations, while the PSSUQ/CSUQ has a somewhat wider range of items, we suggest the SUS if general usability is the focus and/or if a measure with strong statistical support is required. Whereas, if a much wider range of usability aspects and/or specific subscale scores is desired, one of those differentiating between pragmatic and hedonic quality would be recommended. Out of these, the meCUE appears to have a slight advantage in the amount of statistical support, but with the caveat that this is a recently developed questionnaire, so the extent of support does not match some earlier developed measures. For post-task questionnaires, the SEQ or SMEQ are the most straightforward measure of the difficulty in performing a task, along with having good statistical support, so are potentially preferable to other measures.

Future research could expand on this review by exploring the most appropriate questionnaire(s) for a specific set of contexts to further guide questionnaire selection. Multiple questionnaires each assessing different aspects of usability and/or components of a usability study (e.g., a post-task measure combined with a post-study measure) might be the best solution. In addition, for most questionnaires, many aspects of validity and sensitivity are unknown, because they have not been correlated with all usability measures or predicted outcomes, or all the possible relevant sensitivity categories tested. Additional research in this area would provide a clearer general comparison between measures, and further direct comparisons would strengthen this. Lastly, the overlap between usability and workload could be further explored to decide if workload measures would complement a usability study.

Conclusion

As outlined, there are many subjective measures of usability to choose from when designing a usability study. This review highlights that the aims of the study, the system to be assessed, and the specific context need to be considered. This should help guide which type of questionnaire should be used (or in some situations both types) to then compare between multiple questionnaires. The latter comparison should consider the general content, respective advantages and disadvantages, coverage of usability aspects, and psychometric support when choosing the most appropriate measure(s) to use. The review highlights such aspects and, where appropriate, weighs up questionnaires on a specific aspect (so practitioners can identify their relative strengths) along with providing general questionnaire recommendations.

Tips for Usability Practitioners

This review paper highlights the potential confusion over which usability methods are appropriate to use for a given usability study, specifically which general questionnaire(s) to use. Practitioners can use the following tips when determining which questionnaire to use for their study:

- In choosing an appropriate questionnaire to measure usability, practitioners should first identify the relevant category to select questionnaires from. This can be identified through consideration of the evaluation study context, along with the specific system under evaluation.
- Once the most appropriate category has been identified, practitioners should consider the questionnaire aspects that are most relevant and/or important to their study and system. These should include general structure and content details (including the wording of individual items), specific advantages or disadvantages, aspects of usability covered, and psychometric support.
- Practitioners should consider which questionnaires are strongest in assessing usability, and/or cover specific research needs, in each of the relevant areas. For example, it might be considered important to use a measure that displays a range of evidence of validity or more specifically has criterion validity from studies finding a correlation between the measure and another specific measure.
- Practitioners should consider which questionnaires cover specific research needs. For example, questionnaires that are short and quick to complete or freely available may be considered important.

Acknowledgments

The authors express their gratitude to the Global Challenges Research Fund (GCRF) and the Engineering and Physical Sciences Research Council (EPSRC) for the financial support under the International Grant, EP/PO28543/1, entitled "A Collaborative Multi-Agency Platform for Building Resilient Communities." Thanks also to Hanneke Van-Dijk for reviewing the paper.

References

Albert, W., & Dixon, E. (2003, June). Is this what you expected? The use of expectation measures in usability testing. In *Proceedings of Usability Professionals Association 2003 Conference, Scottsdale, AZ*.

Albert, W., & Tullis, T. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Morgan Kaufmann.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594.

Bangor, A., Kortum, P. & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Benedek, J., & Miner, T. (2002). Measuring Desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, *2003*(8–12), 57.

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, *31*(8), 484–495.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988, May). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 213–218).

Cordes, R. E. (1984, October). Software ease-of-use evaluation using magnitude estimation. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 28, No. 2, pp. 157–160). SAGE Publications.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *(13)*3, 319–340.

De La Cruz, J. C. A. (n. d.). Application of Purdue Usability Testing Questionnaire (PUTQ) in the assessment of the Human Computer Interface Factors that measures Intelligence Index of Interface Usability. https://www.academia.edu/26668644/Application_of_Purdue_Usability_Testing_Questionnaire_PUTQ_in_the_assessment_of_the_Human_Computer_Interface_Factors_that_measures_Intelligence_Index_of_Interface_Usability?form=PUTQ

Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, *12*(2), 259–274.

Doll, W. J., Xia, W., & Torkzadeh, G. (1994). A confirmatory factor analysis of the end-user computing satisfaction instrument. *MIS Quarterly*, *(18)*4, 453–461.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000, April). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 345–352).

Gao, M., Kortum, P., & Oswald, F. (2018, September). Psychometric evaluation of the USE (Usefulness, Satisfaction, and Ease of use) questionnaire for reliability and validity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 62, No. 1, pp. 1414–1418). SAGE Publications.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*. http://hdl.handle.net/1805/344

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 9, pp. 904–908). SAGE publications.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction*, *19*(4), 319–349.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187–196). Vieweg+ Teubner Verlag.

Hassenzahl, M., & Sandweg, N. (2004, April). From mental effort to perceived usability: Transforming experiences into summary assessments. In *CHI'04 extended abstracts on Human Factors in Computing Systems* (pp. 1283–1286).

Isleifsdottir, J., & Larusdottir, M. (2008, June). Measuring the user experience of a task oriented software. In *Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement* (Vol. 8, pp. 97–101).

ISO (1998). ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability.

Kirakowski, J. (n. d.). SUMI. *The de facto industry standard evaluation questionnaire for assessing quality of use of software by end users*. http://sumi.uxp.ie/index.html

Kirakowski, J. (1994). The use of questionnaire methods for usability assessment. *Unpublished manuscript. Recuperado el*, *12*.

Kirakowski, J., & Cierlik, B. (1998, October). Measuring the usability of web sites. In *Proceedings of the Human Factors and Ergonomics Society annual meeting* (Vol. 42, No. 4, pp. 424–428). SAGE Publications.

Kortum, P., & Johnson, M. (2013, September). The relationship between levels of user experience with a product and perceived system usability. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 197–201). SAGE Publications.

Lah, U., Lewis, J. R., & Šumak, B. (2020). Perceived usability and the modified Technology Acceptance Model. *International Journal of Human–Computer Interaction*, *(36)*13, 1–15.

Laugwitz, B., Held, T., & Schrepp, M. (2008, November). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group* (pp. 63–76). Springer-Verlag.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *ACM SIGCHI Bulletin*, *23*(1), 78–81.

Lewis, J. R. (1992, October). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 36, No. 16, pp. 1259–1260). SAGE Publications.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use*. International Journal of Human-Computer Interaction*, *(7)*1, 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*(3–4), 463–488.

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction*, *34*(12), 1148–1156.

Lewis, J. R. (2019). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human–Computer Interaction*, *35*(15), 1404–1419.

Lewis, J. R., & Sauro, J. (2009, July). The factor structure of the system usability scale. In *International Conference on Human Centered Design* (pp. 94–103). Springer-Verlag.

Lewis, J., & Sauro, J. (2017). Revisiting the factor structure of the System Usability Scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lewis, J., & Sauro, J. (2020). *Three branches of standardized UX measurement*. Measuring U. https://measuringu.com/three-branches-ux/

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2102).

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505.

Lin, H. X., Choong, Y. Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4-5), 267–277.

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, *13*(8). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199661

Lund, A. M. (2001). Measuring usability with the use questionnaire. *Usability Interface*, *8*(2), 3–6.

Madan, A., & Dubey, S. K. (2012). Usability evaluation methods: A literature review. *International Journal of Engineering Science and Technology*, *4*(2), 590–599.

McGee, M. (2003, October). Usability magnitude estimation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, No. 4, pp. 691–695). SAGE Publications.

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on System Usability Scale ratings. *Journal of Usability Studies*, *7*(2), 56–67.

Minge, M., Thüring, M., & Wagner, I. (2016, September). Developing and validating an English version of the meCUE questionnaire for measuring user experience. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 2063–2067). SAGE Publications.

Minge, M., Thüring, M., Wagner, I., & Kuhr, C. V. (2017). The meCUE questionnaire: A modular tool for measuring user experience. In *Advances in Ergonomics Modeling, Usability & Special Populations* (pp. 115–128). Springer.

Peres, S. C., Pham, T., & Phillips, R. (2013, September). Validation of the system usability scale (SUS) SUS in the wild. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 192–196). SAGE Publications.

Sagar, K., & Saha, A. (2017). A systematic review of software usability studies. *International Journal of Information Technology*, 1–24.

Sauro, J. (2011a). *A practical guide to the System Usability Scale: Background, benchmarks, & best practices.* Measuring Usability, LLC.

Sauro, J. (2011b). *Measuring usability with the system usability scale (SUS)*. Measuring U. https://measuringu.com/sus/

Sauro, J. (2011c). *Does prior experience affect perceptions of usability?* Measuring U. https://measuringu.com/prior-exposure/

Sauro, J. (2012). *10 things to know about the Single Ease Question (SEQ)*. Measuring U. https://measuringu.com/tag/seq/

Sauro, J., & Dumas, J. S. (2009, April). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1599–1608).

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014, June). Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In *International Conference of Design, User Experience, and Usability* (pp. 383–392). Springer.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *IJIMAI*, *4*(6), 103–108.

Shneiderman, B. (1987). *Designing the user interface: Strategies for effective human-computer interaction*. Addison-Wesley Publishing Co.

Tedesco, D., & Tullis, T. (2006). A comparison of methods for eliciting post-task subjective ratings in usability testing. In *Usability Professionals Association (UPA) Conference* (pp. 1–9).

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, *42*(4), 253–264.

Tullis, T. S., & Stetson, J. N. (2004, June). A comparison of questionnaires for assessing website usability. In *Usability Professional Association (UPA) Conference* (Vol. 1).

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, (26)*2, xiii–xxiii.

Wohlin, C. (2014, May). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (pp. 1–10).

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools* [Doctoral dissertation, Delft University of Technology]. TU Delft Research Repository. https://repository.tudelft.nl/islandora/object/uuid%3Ad97a028b-c3dc-4930-b2ab-a7877993a17f

About the Authors

Andrew Hodrien
Dr. Hodrien is a Psychologist who received his PhD in Applied Psychology at the University of Salford. His PhD focused on user experience of prosthetics, with recent research (THINKlab, University of Salford) focusing on assessment of an immersive VR crowd-control simulation and usability of a multi-agency collaboration platform (MOBILISE project).

Terrence Fernando
Professor Fernando (Director of the THINKlab at the University of Salford) is experienced in conducting multi-disciplinary/international research programs. His research focuses on disaster management, collaborative working environments, building simulation, urban simulation, and smart cities.